Cognitive Model Priors for Predicting Human Decisions

David Bourgin*[1] Joshua Peterson*[2] Daniel Reichman[2] Stuart Russell[1] Thomas Griffiths[2]

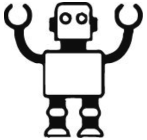
[1]University of California, Berkeley, [2]Princeton University

ICML 2019

Predicting **human behavior** is important for…

Economics

Psychology



AI-human Alignment

Two Approaches

Two Approaches

Behavioral Science

Step 1
Observe behavior

Step 2
Create theory / model

$$\sum_{j=1}^{N} p_j u(r_j)$$

$$\sum_{j=1}^{N} \pi(p_j) v(r_j)$$

Two Approaches

Behavioral Science

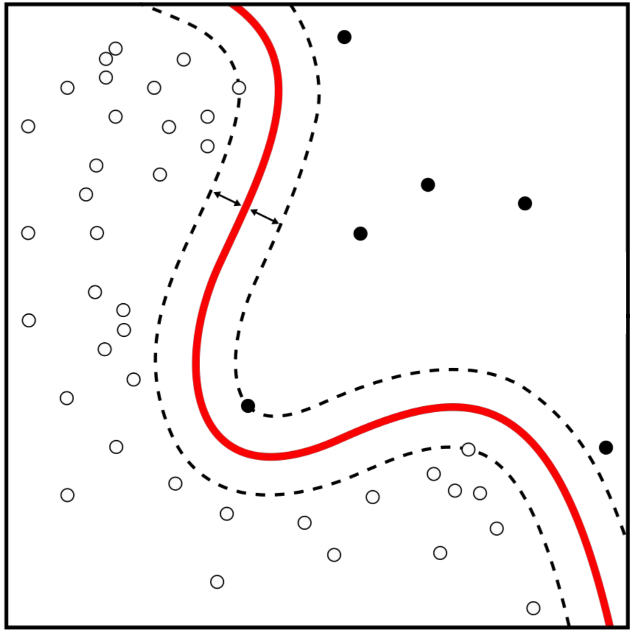
Step 1
Observe behavior

Step 2
Create theory / model

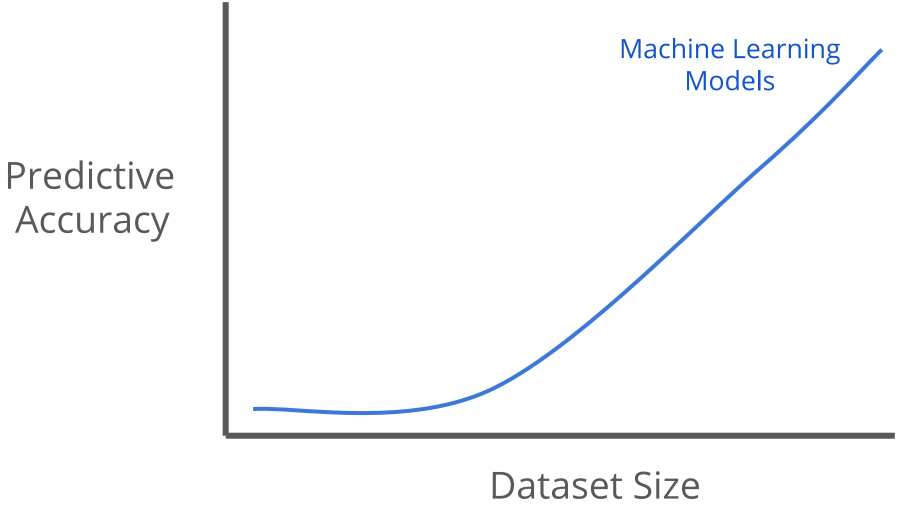
$$\sum_{j=1}^{N} p_j u(r_j)$$

$$\sum_{j=1}^{N} \pi(p_j) v(r_j)$$

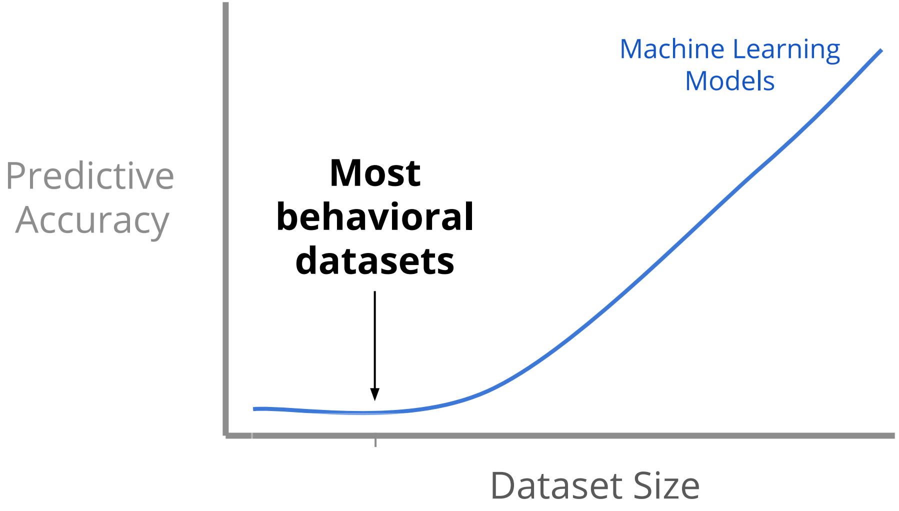
Machine Learning



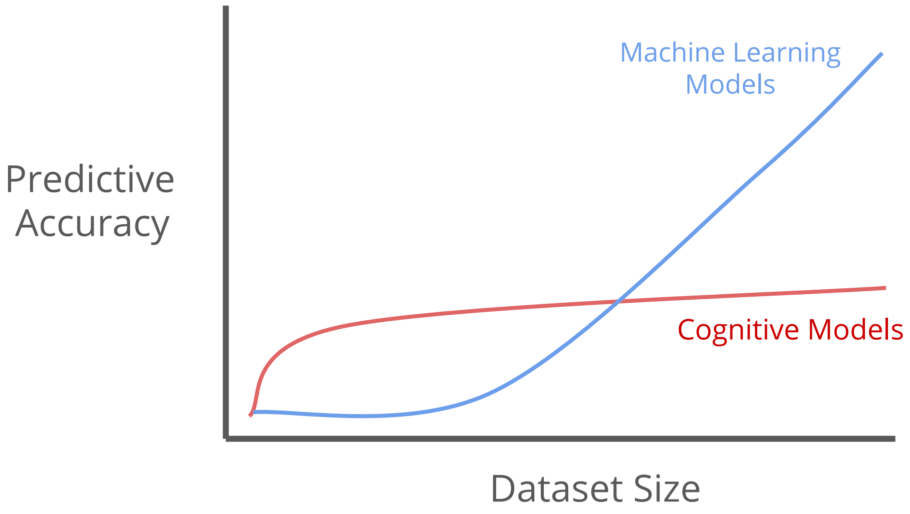
ML can be very effective, but **needs lots of data**



ML can be very effective, but **needs lots of data**



ML can be very effective, but **needs lots of data**



Cognitive models need less data, but **improve slower**

Cognitive Model Priors

Cognitive Model Priors

1. Use a cognitive model to generate **synthetic behavioral data**

Cognitive Model Priors

1. Use a cognitive model to generate **synthetic behavioral data**

2. **Pretrain** a neural network on this synthetic behavior

Cognitive Model Priors

1. Use a cognitive model to generate **synthetic behavioral data**

2. **Pretrain** a neural network on this synthetic behavior

3. **Fine-tune** the pretrained network on real human behavior

Case Study: Risky Choice

- Choices that involve uncertainty & monetary gain/loss
- Multiple models developed over decades

Kahneman & Tversky (1979)
Peysakhovich et al. (2017)
Erev et al. (2017)

Task is to **choose between two gambles**

A

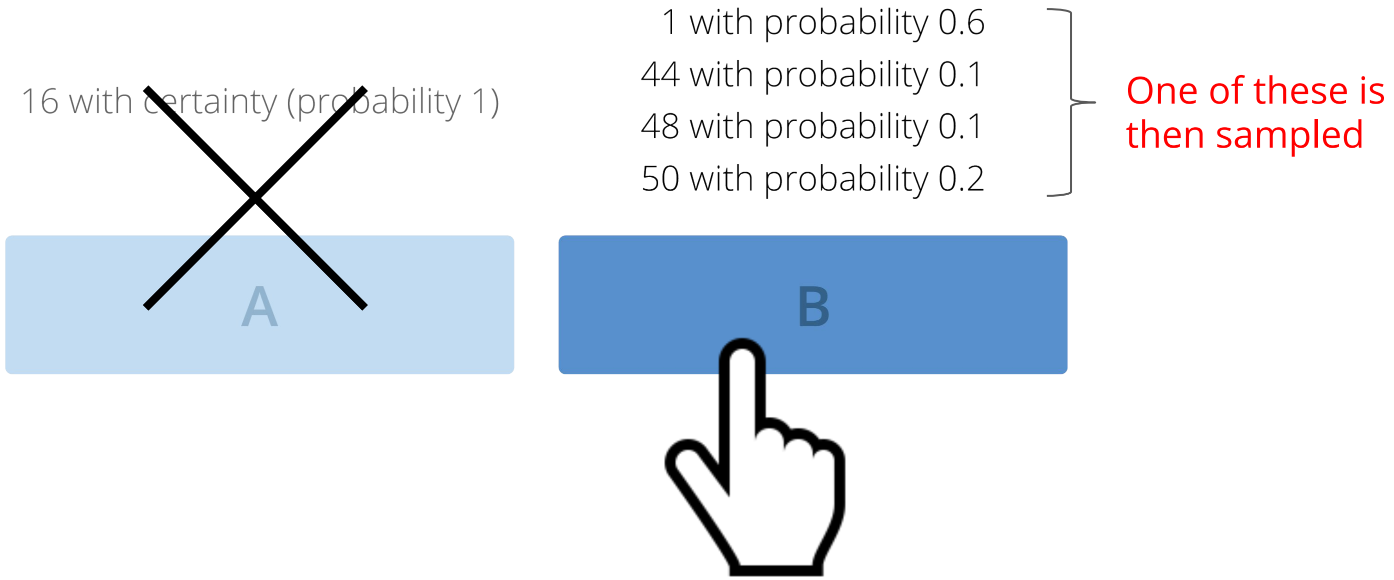
B

A **gamble** is a collection of outcomes (*rewards*) & their probabilities

16 with certainty (probability 1)

1 with probability 0.6
44 with probability 0.1
48 with probability 0.1
50 with probability 0.2

A	B


16 with certainty (probability 1)
A
1 with probability 0.6
44 with probability 0.1
48 with probability 0.1
50 with probability 0.2
B
One of these is then sampled

1 with probability 0.6
44 with probability 0.1
48 with probability 0.1
50 with probability 0.2

16 with certainty (probability 1)

A	B

Feedback: You chose **B** and gained **50**

Had you chosen A, you would have gained 16

Cognitive Models of “risky” decision-making

(between gambles)

Cognitive Models of “risky” decision-making

(between gambles)

Approach

1. Specify the **subjective value** of a gamble
2. Choose gamble with **highest value**

Cognitive Models of “risky” decision-making

(between gambles)

Approach

1. Specify the **subjective value** of a gamble
2. Choose gamble with **highest value**

Lots of models we could use…

$$\sum_{j=1}^{N} p_j r_j$$

$$\sum_{j=1}^{N} p_j u(r_j)$$

$$\sum_{j=1}^{N} \pi(p_j) v(r_j)$$

…

Cognitive Models of “risky” decision-making

(between gambles)

Approach

1. Specify the **subjective value** of a gamble
2. Choose gamble with **highest value**

Lots of models we could use...

$$\sum_{j=1}^{N} p_j r_j$$

$$\sum_{j=1}^{N} p_j u(r_j)$$

$$\sum_{j=1}^{N} \pi(p_j) v(r_j)$$

...

We used SOTA: “**BEAST**”

- Estimates expected value (payoff) with biased, sampled-based, estimators
- **We treat as black box with inputs/outputs**

Erev et al.. *Psychol. Rev.*, 2017, *124*, 369.
Plonsky et al. 2019, arXiv preprint arXiv:1904.06866.

	Model	MSE×100
CPC 2015		
CPC 2018		

CPC15 and **CPC18** competition datasets are still **small** by ML standards

	Model	MSE×100
CPC 2015	*ML + Raw Data*	
	MLP	7.39
	k-Nearest Neighbors	7.15
	Kernel SVM	5.52
	Random Forest	6.13
CPC 2018		

Machine learning struggles when learning from raw inputs and **scarce data**

	Model	MSE×100
CPC 2015	*ML + Raw Data*	
	MLP	7.39
	k-Nearest Neighbors	7.15
	Kernel SVM	5.52
	Random Forest	6.13
	Theoretical Models	
	BEAST15	0.99
	CPC 2015 Winner	0.88
CPC 2018	*Theoretical Models*	
	BEAST18	0.70

Hand-built **cognitive models** do much better

	Model	MSE×100
CPC 2015	*ML + Raw Data*	
	MLP	7.39
	k-Nearest Neighbors	7.15
	Kernel SVM	5.52
	Random Forest	6.13
	Theoretical Models	
	BEAST15	0.99
	CPC 2015 Winner	0.88
	ML + Feature Engineering	
	MLP	1.81
	k-Nearest Neighbors	1.62
	Kernel SVM	1.01
	Random Forest	0.87
	Ensemble	0.70
CPC 2018	*Theoretical Models*	
	BEAST18	0.70
	ML + Feature Engineering	
	Random Forest	0.68
	CPC 2018 Winner	0.57

Machine learning with lots of **feature-engineering** finally shows improvements

2015 winner

Our 2018 winning entry

	Model	MSE×100
CPC 2015	*ML + Raw Data*	
	MLP	7.39
	k-Nearest Neighbors	7.15
	Kernel SVM	5.52
	Random Forest	6.13
	Theoretical Models	
	BEAST15	0.99
	CPC 2015 Winner	0.88
	ML + Feature Engineering	
	MLP	1.81
	k-Nearest Neighbors	1.62
	Kernel SVM	1.01
	Random Forest	0.87
	Ensemble	0.70
	MLP + Cognitive Prior (ours)	**0.53**
CPC 2018	*Theoretical Models*	
	BEAST18	0.70
	ML + Feature Engineering	
	Random Forest	0.68
	CPC 2018 Winner	0.57
	MLP + Cognitive Prior (ours)	**0.48**

Our method
outperforms them all

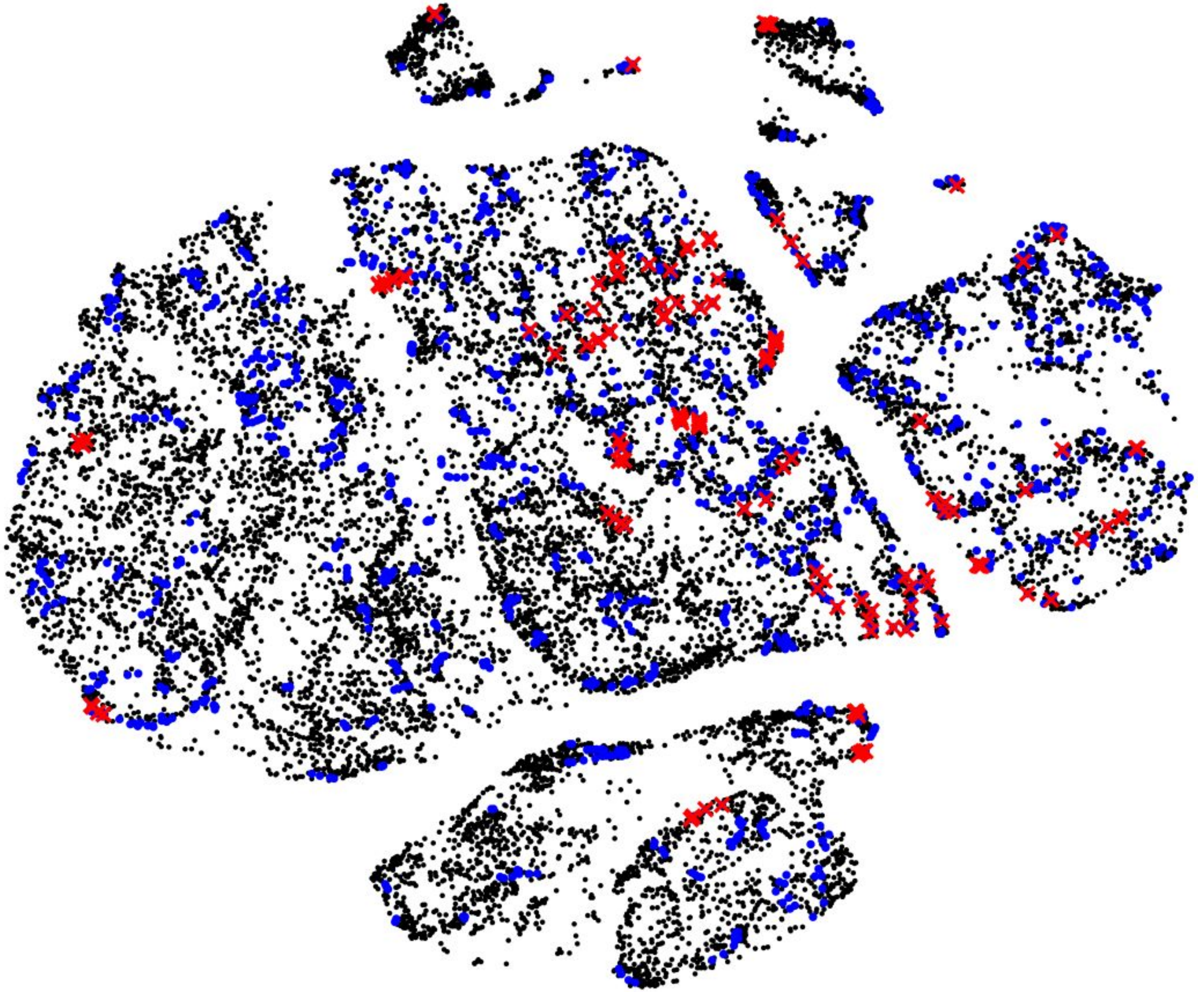
Better than our CPC18 winner

Result: `choices13k` dataset

- 13,000 pairs of gambles
- 240k individual decisions

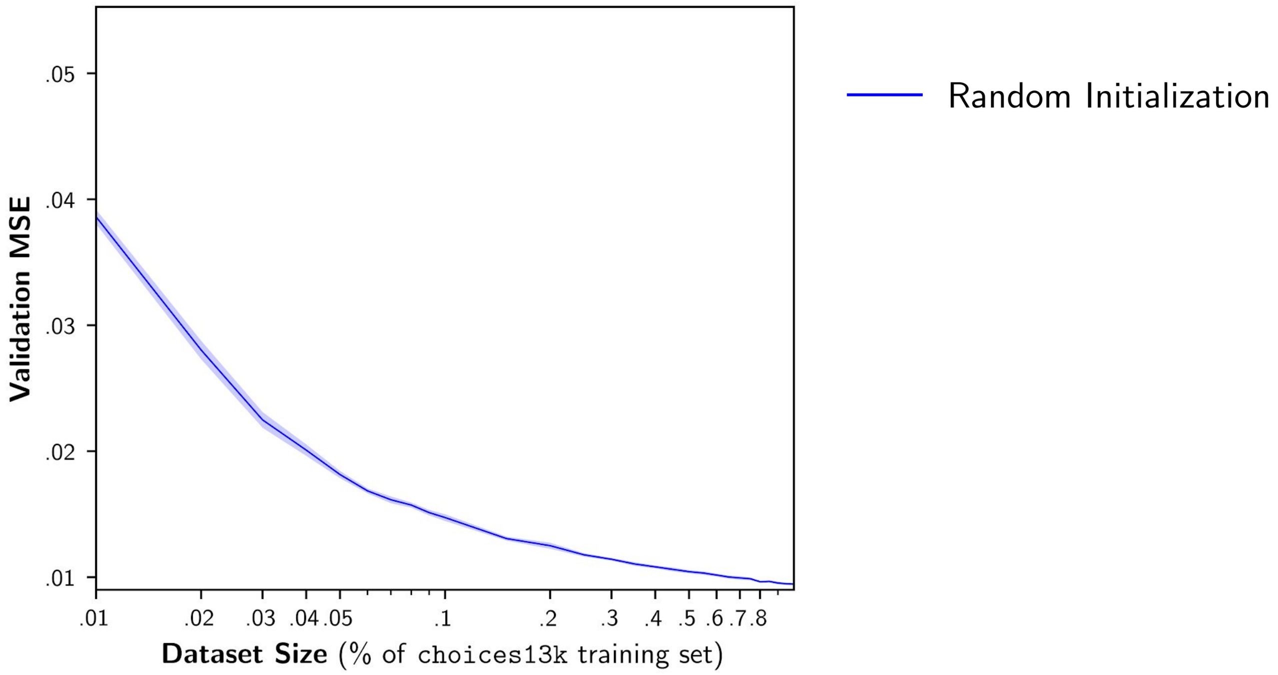
Result: `choices13k` dataset

- 13,000 pairs of gambles
- 240k individual decisions



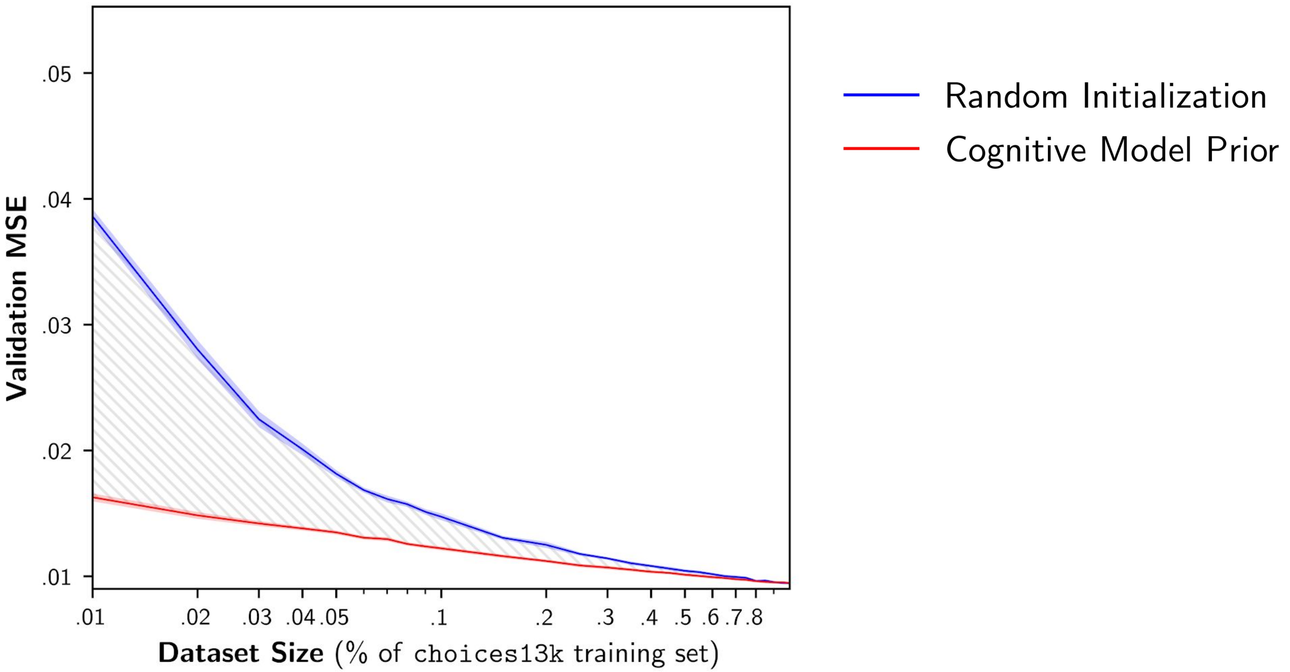
- ❌ Classic Experiments
- ● Previous Benchmark (CPC)
- ● **Ours:** `choices13k`

New dataset lets us compare **different levels of data scarcity**...

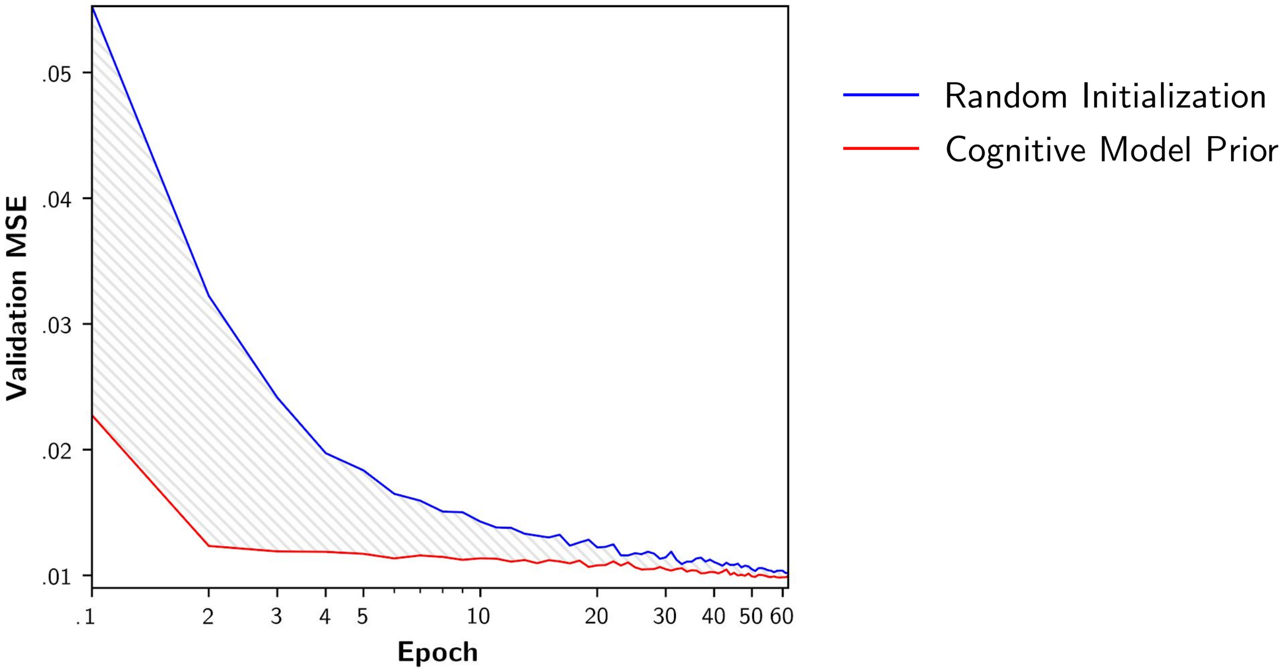


Random Initialization
Validation MSE
.05
.04
.03
.02
.01
.01
.02
.03
.04
.05
.1
.2
.3
.4
.5
.6
.7
.8
Dataset Size (% of choices13k training set)

When data is scarce, cognitive model priors **improve generalization**



When data is scarce, cognitive priors **reduce training time**



Predicting **human behavior** is important for...

- Economics
- Psychology
- AI-Human Alignment

Cognitive model priors **improve accuracy** and **reduce training time**

Cognitive Model Priors for Predicting Human Decisions

David Bourgin* Joshua Peterson* Daniel Reichman Stuart Russell Thomas Griffiths

Co-authors

Joshua Peterson

Daniel Reichman

Stuart Russell

Thomas Griffiths

Funding

DARPA

Future of Life Institute

Open Philanthropy Project

National Science Foundation

Poster **#244**

Wednesday Evening

ddbourgin@gmail.com

peterson.c.joshua@gmail.com